PB86-199049

Computer-Assisted Studies of Molecular Structure and Genotoxic Activity by Pattern Recognition Techniques

Pennsylvania State Univ.

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

1985

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA/600/J-85/384		PB86-199049

4. TITLE AND SUBTITLE	5. REPORT DATE
COMPUTER-ASSISTED STUDIES OF MOLECULAR STRUCTURE AND GENOTOXIC ACTIVITY BY PATTERN RECOGNITION TECHNIQUES (Journal Version)	1985
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Terry R. Stough and P.C. Jurs	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
Pennsylvania State University, University Park, PA 16802	11. CONTRACT/GRANT NO. CR811037
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Health Effects Research Laboratory - RTP, NC Office of Research and Development US Environmental Protection Agency Research Triangle Park, NC 27711	14. SPONSORING AGENCY CODE EPA/600/11

15. SUPPLEMENTARY NOTES

Journal - Environmental Health Perspectives Vol. 61, pp. 329-343, September 1985.

16. ABSTRACT

Often a compound's biological activity is determined by complex relationships between its structural components. Such a relationship often can only be adequately described and exploited by multivariate structure-activity relationship (SAR) studies that can deal with many variables simultaneously. Pattern recognition (PR) is a multivariate technique that is well suited for the qualitative, active-inactive, data that is often supplied by biological assays. PR studies of compounds of known activity can yield information that will allow the prediction of the activity of untested compounds. ADAPT is a computerized system that was developed for such PR-SAR studies. A general introduction to this field is presented and the methodology used for such a study is described in the context of an actual study of mutagenic compounds. The data requirements, descriptor generation, and the details of a PR study are discussed. In addition, the example study was chosen to highlight the problems that may occur if a study is not well formulated and carefully executed. Current work and future plans for computerized mutagen screening are discussed.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field Group

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	16
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

Environmental Health Perspectives
Vol. 61, pp. 329-343, 1985

EPA/600/J-85/384
JOURNAL ARTICLE

Computer-Assisted Studies of Molecular Structure and Genotoxic Activity by Pattern Recognition Techniques

by Terry R. Stouch* and Peter C. Jurs*†

Often a compound's biological activity is determined by complex relationships between its structural components. Such a relationship often can only be adequately described and exploited by multivariate structure–activity relationship (SAR) studies that can deal with many variables simultaneously. Pattern recognition (PR) is a multivariate technique that is well suited for the qualitative, active-inactive, data that is often supplied by biological assays. PR studies of compounds of known activity can yield information that will allow the prediction of the activity of untested compounds. ADAPT is a computerized system that was developed for such PR–SAR studies. A general introduction to this field is presented and the methodology used for such a study is described in the context of an actual study of mutagenic compounds. The data requirements, descriptor generation, and the details of a PR study are discussed. In addition, the example study was chosen to highlight the problems that may occur if a study is not well formulated and carefully executed. Current work and future plans for computerized mutagen screening are discussed.

Introduction

Pattern recognition (PR), a subfield of artificial intelligence, consists of several loosely related mathematical techniques. The goal of these techniques is to classify numerical patterns into one of several possible classes. In structure–activity relationship (SAR) studies, the patterns consist of vectors of 'measurements' made of the compounds in the study. Each class consists of compounds of like activity. A PR–SAR analysis entails finding those structural features that will define the distinct classes of activity. This is often performed first for a training set of compounds of known activity. The results for the training set can then be applied to predict the activity of unknowns.

PR techniques fall into several categories: cluster analysis (*1*), mapping (*2*), discriminant generation (*3*), and principal components analysis (*4*). Several computerized systems have been developed for conducting PR studies: ADAPT (*5*), ARTHUR (*6*), and SIMCA (*4*). Many SAR–PR studies have appeared in the literature (*7–28*), and several reviews and books deal in part or in whole with such studies (*5,10,18,19*). SIMCA will be discussed in detail elsewhere in this volume. Our intent is not to duplicate these reviews but to clarify, by example, the methodology and problems involved in studies applying linear discriminant functions (LDFs).

Our research group has been involved in the development and use of the interactive software system ADAPT (Automated Data Analysis using Pattern Recognition Techniques). This system conveniently integrates the many steps required for a complete SAR study not only using PR but also using a variety of statistical techniques. Modular routines provide for structure entry, descriptor generation, and data analysis. ADAPT studies have been performed on antitumor agents (*11*), carcinogenic substances (*8,9,12,13*), and olfactory stimulants (*10*).

In this paper, the methodology used for such a study is described in the context of an actual study of mutagenic compounds. The data requirements, descriptor generation and the details of a PR study are discussed. In addition, the example study was chosen to highlight the problems that can occur if a study is not well formulated and carefully executed.

Theory

A discussion of theory will help to clarify the rest of this paper. Very simply put, the goal of an SAR study is to find a function of structural or physical properties that will explain a compound's activity relative to other similar compounds. A PR–SAR study is based on sev-

*Department of Chemistry, 152 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802.
†Author to whom reprint requests should be sent.

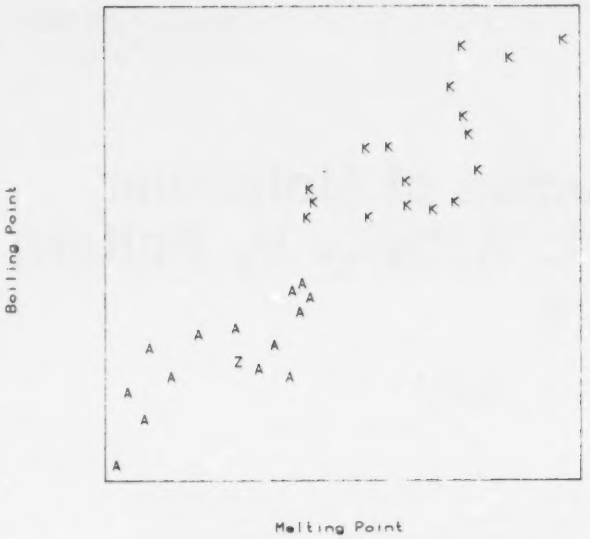


FIGURE 1. Boiling point vs. melting point for some simple aldehydes (A) and ketones (K). Note that the two classes cluster in separate regions of the plot. Z is a compound of unknown classification.

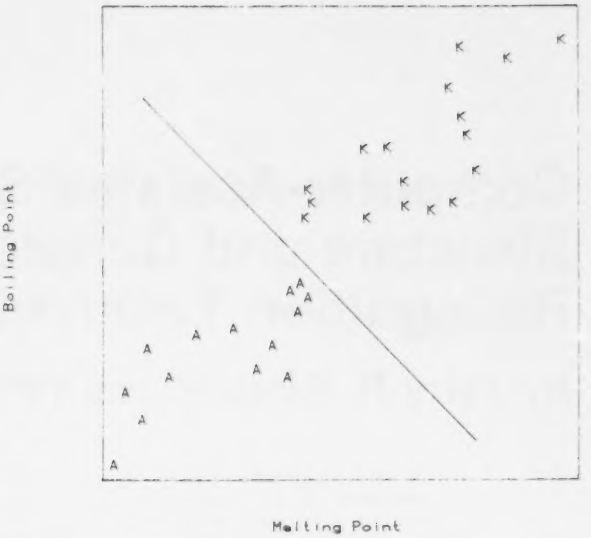


FIGURE 2. A one-dimensional hyperplane that separates the two regions which contain the two different classes.

eral assumptions: (1) The activity and its variation can be explained by variations within the structures. (2) The structures can be sufficiently described by numerical indices (descriptors). (3) Pattern recognition techniques can be used to discover a relationship between the descriptors and the activity. (4) This relationship can be extrapolated to untested compounds.

Each compound in a study is referred to as an observation or pattern and each structural feature or experimental property is referred to as a variable or descriptor. Simple problems can be viewed graphically as in Figure 1. This is a plot of several aldehydes (A) and ketones (K) as represented by their melting points and boiling points. In the "space" of these two physical parameters, the aldehydes cluster in a different region of this "two-space" than the ketones. This differential clustering is the basis for pattern recognition. If a new compound (Z) is plotted in this same space, the likelihood is high that it will belong to the same class as neighboring patterns, in this case, aldehydes. This is a very simple example of cluster analysis. Techniques that generate discriminant functions also rely on this clustering, but instead of comparing an observation to others they simplify the problem by generating a boundary that will separate the regions of the space which contain the two clusters (Fig. 2). In this two-space the boundary consists of a line. In three dimensions, a plane would be used, and in higher dimensions, a hyperplane. We are most concerned with these higher dimensional problems. Simple one- and two-dimensional problems can be approached visually, but higher dimensional problems require the use of computer-aided mathematics. It should be noted that the boundary shown in Figure 2 is only one of the many which will separate the two classes.

LDFs are often referred to as weight vectors. An LDF is a vector of weights (coefficients) for the descriptors. A plane, say the boundary between two classes, is defined by a point on that plane and a vector normal to the plane's surface. Often an extra dimension is added to the data so that the plane includes the point at the origin (O,O,...,O). It is this point and the weight vector that define the boundary. Adding the extra dimension in this way greatly simplifies the calculations.

The discriminant lines, planes, and hyperplanes are defined by linear combinations of the descriptors used in their development and so the weight vectors are referred to as linear discriminant functions. While nonlinear discriminants can also be used, they are more difficult to deal with both conceptually and computationally. Also, many nonlinear relationships can be investigated by scaling or transforming the variables. Linear SAR studies routinely use the logarithm of a variable in order to linearize exponentials. For these reasons, studies using linear discriminant functions are more common than those using nonlinear functions, and we will restrict our discussions to the linear case. Scaling and normalization will not be discussed in detail here; the interested reader can refer to the literature (*3*,*5*).

There are many ways to generate LDFs. Parametric methods calculate the LDF using class statistics that are based on some assumed distribution, usually a multivariate normal distribution. Nonparametric methods use the information contained in the individual patterns to determine the placement of the hyperplane. Nonparametric methods include the linear learning machine (*20*) and several least-squares methods (*21*,*22*) as well as others. Nonparametric methods are applied to problems for which the distribution is unknown and difficult to determine.

Regression analysis is a widely used technique in SAR. Since it bears some superficial resemblances to PR, and since some tend to confuse the capabilities of the two, we would like to draw a contrast between these methods. Both use numerically encoded structural features or physical properties to represent the compounds, and both generate a vector of weights that describe a relationship between those descriptors and the activities of the compounds. The similarity ends at this point. The data requirements, capabilities, and weight vectors are very different.

Regression analysis finds that line, plane, or hyperplane that best "fits" the data. PR analysis, as was noted previously, finds a hyperplane that separates the data points into two or more regions of the data space. Regression analysis determines a unique hyperplane by finding those weights that minimize a function of the difference between the actual and predicted activity. The hyperplane represented by the LDF is not unique and is determined by the general regions of the space in which the classes fall. Given the same data, a plane represented by an LDF might lie at a 90° angle to the regression-generated plane.

PR and regression analysis are best suited to different types of problems. The capabilities of regression analysis are best realized when a quantitative index of activity is provided. Such information is often difficult or expensive to obtain. Often qualitative, active–inactive, information is more readily available and such data is better suited to the classification routines of PR. The type of data supplied for such a study, quantitative or qualitative, determines the type of information which a study can supply in return. An analysis cannot generate new data and cannot supply quantitative information when only qualitative information is provided. Regression studies are often used to supply a quantitative prediction of the activity of an unknown. This is reasonable if the activities of the compounds used to generate the regression model were also known quantitatively. PR, however, is usually provided only with qualitative, class information and no quantitative predictions can be expected in return. Additionally, since the LDFs are nonunique, any quantitative index will vary depending on which of these vectors is used.

A related topic is that of the interpretation of the coefficients of the regression model or the LDF. The nature of regression analysis allows for evaluation of the importance of a particular descriptor to a model by examination of its partial F-statistic. In fact, variables are chosen for their ability to explain the variance in the activity. Once again, less information is provided to PR routines and so less should be expected in return. Coefficients of an LDF should be interpreted with care. Descriptor importance is dependent on the complete set of descriptors used. The overall set is chosen for its ability to discriminate between the classes and an individual variable's contribution to the set may vary from its dicriminatory ability alone.

Finally, a few words about the number of observations required to conduct a PR study. There are no firm guidelines to specify this parameter. Depending on the type of structures, the activity which determines the classifications, and the descriptors employed, the number of compounds needed for a study will vary. Most ADAPT studies have been conducted with well over 50 compounds; many range into the hundreds. The more information that is provided to the PR routines, the more reliable will be the results. Also, as will soon be explained, the number of patterns that are involved in a study determines the allowable number of descriptors that can be used to describe those patterns. Small studies must necessarily be 'tight' structurally or mechanistically because there will only be a few allowable features with which to describe them.

This brief discussion of some aspects of PR and regression analysis has introduced some ideas to be used in the remainder of this paper. More complete discussions of these methods can be found in the literature (*3*,*5*,*20*–*26*).

Data

The first step of an analysis is crucial—development of the data set which will be investigated. From the compounds in this set, the training set, important descriptors will be identified from which the LDFs will be generated. The only information which the PR routines have to work with is provided by this training set as it is represented by the descriptors. The predictive ability of an analysis will depend on the accuracy, completeness, and sufficiency of the data. Later in this paper, the choice of data will be discussed further.

We have been working with the Metabolism and Carcinogenesis Branch of the U.S. Environmental Protection Agency towards developing a computerized method of screening potential mutagens. Thousands of chemicals are submitted for regulatory decisions each year and it is impossible both financially and practically to perform full-scale laboratory testing on them all. A computerized system that could readily identify potential mutagens would be valuable in prioritizing the laboratory testing.

The initial goal was to conduct SAR studies that would duplicate some *in vitro* assays that provide valuable screening information. The assay that we have examined most completely is the *E. coli* WP2uvrA reverse-mutation assay. This assay was of interest because it has been used widely and is well documented (*27*). Also, a set of compounds screened in this assay were available through a published report of the Gene-Tox program (*27*,*28*). This program is responsible for evaluating genetic toxicity assays. It often publishes lists of compounds that have been tested in the assays and that the authors, a panel of experts, feel have been properly assayed and classified.

The report on the *E. coli* WP2uvrA tryptophan reversion assay contained a list of 158 compounds that had been assayed using this and another *E. coli* assay. Of

Table 1. Training set compounds.

No.	Compounds	No.	Compounds
Mutagens		Nonmutagens	
1	Nifuroxime	51	Benzoic acid
2	Nitrofurantoin	52	Quintozene
3	Captafol	53	Furaldehyde semicarbazone
4	Dimethylnitrosamine	54	2,4-Dichlorophenoxy-butanoic acid
5	5-Nitro-2-furanoic acid	55	1,3,5-Trichloro-2,4,6-trinitrobenzene
6	Dichlorvos	56	Acephate
7	Henomyl	57	Aspon
8	Mitomycin-C	58	Azinphos-methyl
9	Methylhydrazine	59	Mannitol
10	Misonidazole	60	Carbofuran
11	NF-67	61	Dicamba
12	Nifuratrone	62	Disulfoton
13	Trichlorfon	63	Monocrotaline
14	Styrene oxide	64	Ethion
15	*N*-Methyl-*N'*-nitronitrosoguanine	65	Fonofos
16	Hydroxyethylhydrazine	66	Fensulfothion
17	Cyclophosphamide	67	Malathion
18	7-Bromomethylbenz(a)anthracene	68	Parathion-methyl
19	Nitrofluorene	69	Simazine
20	Ethylmethanesulfonate	70	Trifluralin
21	Metronidazole	71	*N*-[4[(5-Nitro-2-furyl)-2-thiazolyl]acetamide
22	Niridazole	72	2,2,2-Trifluoro-*N*-[4-(5-nitro-2-furyl)-2-thiazolyl]acetamide
23	Dexon		
24	Nitrosomethylurea	73	2,4-Dichlorophenoxyacetic acid
25	Acrylonitrile	74	Adipic acid
26	*p*-Bromostryrene oxide	75	1,3,5-Trichlorobenzene
27	*m*-Chlorostyrene oxide	76	1,3,5-Triamino-2,4,6-trinitrobenzene
28	3,4-Dimethylstyrene oxide	77	*m*-Methoxystyrene oxide
29	*m*-Methylstyrene oxide	78	*N*-[4-(5-Nitro-2-furyl)-2-thiazolyl]phenylamine
30	*p*-Methylstyrene oxide	79	2-Amino-4-(5-nitro-2-furyl)thiazole
31	Sulfur mustard	80	2-(2,2-Dimethylhydrazino)-4-(5-nitro-2-furyl)thiazole
32	Methyl methanesulfonate	81	2,4-Dinitrophenylthiocyanate
33	2-Chloro 4-(5-nitro-2-furyl)thiazole	82	Ethylenethiourea 2-furyl)-vinyl-1,2,4-triazine
34	2-Hydrazino-4-(5-nitro-2-furyl)thiazole	83	5-Nitronaphthonitrile
35	3-Amino-6-[2-(5-nitro-2-furyl)-vinyl-1,2,4-triazine	84	Hippuric acid 2-thiazolyl]-formhydrazide
36	N4-Hydroxycytidine	85	2-[4-(2-Furyl)-2-thiazolyl]formhydrazide
37	*N*-[-(5-Nitro-2-furyl)-2-thaizolyl]formamide	86	2-[4-Methyl-2-thiazolyl] formhydrazide
38	5-Nitro-2-furanmethandiol diacetate	87	O.S--Dibenzoylthiamine HCl
39	Furazolidone	88	Carteoloc HCl
40	4-Nitroquinoline-*N*-oxide	89	Ethylidine gyromitrin
41	2-[4-(5-Nitro-2-furyl)-2-thiazolyl]-formhydrazide	90	1-[[3-(5-Nitro-2-furanyl)-2-propenylindene]amino]-2,4-imidazolidinedione
42	*N*-[3-(5-Nitro-2-furyl)-6H-1,2,4-oxadiazin-5-yl]-acetamide	91	Crotoxyphos
		92	Monocrotophos
43	3-(5-Nitro-2-furyl)-4H-1,2,4-triazole	93	Bromacil
44	1-(5-Nitro-2-furyl)-3-piperidino-propan-1-one semicarbazone HCl	94	Chlorpyrifos
		95	Diazinon
45	Furylfuramide	96	Dinaseb
46	Vamidothion	97	Endrin
47	Furmethanol	98	Fenthion
48	1-(5-Nitro-2-furfurylidine)-3-*N*,*N*-diethyl-propylaminourea HCl	99	Methomyl
		100	Methoxychlor
49	Captan	101	Monuron
50	Folpet	102	Parathion
		103	Phorate
		104	Propanil
		105	Siduron

these, 19 were inorganics, which are not compatible with the ADAPT system or mixtures, which are inappropriate for such a study. Of the 139 compounds remaining, 105 were definite mutagens (*50*) or nonmutagens (*55*) in the assay of interest (Table 1).

These 105 compounds were structurally diverse and ranged in complexity from methylhydrazine to mitomycin-C. Halogenated hydrocarbons, phosphates and thiophosphates, sulfates, polycyclic aromatic hydrocarbons, nitrofurans, substituted benzenes, styrene oxides, and triazines were represented, as well as others. No single type dominated, however, and the largest class had fewer than 10 members.

Normally, SAR studies deal with structurally similar compounds and are concerned with minor changes in structure that can cause great changes in activity. How-

ever, in cases where activity is determined by some general physical property such as size, shape, or solubility, even structurally diverse compounds might be related. Also, if the activity is mediated by a common intermediate or receptor site, only a portion of the moleculesmay be important; that portion which interacts with the receptor or undergoes conversion to the intermediate. Often in problems of biological activity, the mechanism of action is very complex and might be determined by many different effects operating simultaneously. In such a case, multivariate analysis may be the only recourse to discovering, understanding, or explaining this activity. We had hoped to uncover some "core" of mutagenesis among the training set members that would apply to other, untested mutagens.

Once the training set had been established, the next step in the analysis was to provide the structures to the computer. This was accomplished by using an interactive graphics routine, UDRAW (*29,30*). The compounds were sketched by hand on a CRT terminal and the structural information was automatically stored in files where it was available to the descriptor generating programs.

Descriptor Generation

The goal of descriptor generation is to numerically code those properties that may be correlated with the biological activity. This step reduces the structural complexity of a compound to single numbers. Descriptors must be generated and examined carefully in order to be sure that they contain the intended information. Sometimes the needed information can be lost in the transfer from the chemist's perception of the structure to the single numerical value contained in the descriptor.

Many different physical and structural parameters have proven useful in SAR studies. The value of any one descriptor is determined by the study at hand. Sometimes an activity may depend on simple physical properties such as size, volume, or log P. In other cases, the activity may depend on more complex electronic and steric effects that occur simultaneously at different sites. The ADAPT system has a variety of descriptor generating routines that encode a wide range of structural and physical properties. Many properties of whole molecules can be calculated. The log of the partition coefficient between octanol and water, log P, is a parameter that has been very useful in many SAR studies and that can be calculated by the fragment addition method of Hansch and Leo (*31,32*). Molar refractivity is also commonly used and can be calculated using a similar approach (*33*). Surface area and molecular volume (*34,35*) and moments of inertia (*36*) complete the list of whole molecule physical chemical descriptors that ADAPT programs can calculate.

Simple quantum mechanical indices are available in the ADAPT system from del Re sigma charges (*37*) and simple and extended Huckel calculations. Several studies have used this information.

Molecular connectivity as proposed by Randic (*38*) and extended by Kier and Hall correlate highly with many physical properties and have been used in SAR studies (*39*). Molecular connectivity provides indices of a molecule's size and degree of branching.

Substructural descriptors are valuable in emphasizing reactive sites and pharmacophores. While they can serve as simple indicator variables, they can also be used in tandem with molecular connectivity or other calculations to yield information concerning the environment and reactivity of the substructure (*5*). Substructural analysis is common in SAR studies and has been approached by using a variety of techniques, including PR (*40–43*).

For some studies very simple descriptors may be valuable. Counts of the number of a particular atom or bond type and counts of the number of rings are examples of such descriptors and are often referred to as fragment descriptors. Molecular weight is another simply calculated parameter. If experimental information is available, it too can be used.

Prescreening of Descriptors

Before they are used in an SAR study, descriptors should be prescreened using several simple checks. First, a descriptor that codes for very few compounds should be considered only if there is good reason to think that it could be very valuable. Such descriptors can lead to a fortuitous "partitioning off" of those few compounds which they code for and they may not be valuable in explaining overall trends. Ideally, a descriptor should code for all of the observations. Descriptors usually are not considered if they do not code for at least 10% of the observations in each class.

Another simple check is for collinearities and multicollinearities between the descriptors. Collinearities mean that a descriptor can be expressed as a weighted linear combination of other descriptors. For example, molecular weight is a weighted summation of atom counts. Such collinearities needlessly duplicate information. Also, unduly high collinearities among descriptors cause several matrices that arise in the PR phase of the studies to be singular or nearly singular, causing numerical difficulties when the matrices must be inverted.

A final check involves examination of the standard deviation of the descriptors. This is a measure of the spread of values that a descriptor has through the training set of compounds. A descriptor with zero standard deviation provides no information and simply uses up the allowable degrees of freedom that will be discussed shortly.

Autoscaling is often performed prior to PR analysis. This step simultaneously scales and normalizes each descriptor by subtracting that descriptor's mean and dividing by its standard deviation. This gives each descriptor a mean of zero and a standard deviation of one. While the overall distribution of the descriptor values is unaffected, this step removes any weighting effect due to the unit size of the descriptors. Autoscaling pre-

vents descriptors with large absolute values from occluding the effects of descriptors whose absolute values are small in comparison.

Feature Selection

Once a pool of suitable descriptors is available, then how is the first set formulated for PR analysis? It is tempting simply to use very large pools of descriptors hoping first to identify some relationship within that large pool and then progressively to remove those descriptors that contribute little or nothing to that relationship. This approach has several drawbacks. First, Stuper et al. (*44*) have shown that overdescribed cases can actually occlude a separable set. Second, they also determined that feature selection procedures were of little use in identifying useless descriptors in such cases. A more important drawback deals with fortuitous separations. It is well known that for a fixed number of observations the probability of achieving a fortuitous, and probably useless, linear separation increases as the number of descriptors increases (*1*,*3*,*5*,*20*). In fact, if the number of variables is greater than or equal to the number of observations, then 100% separation is guaranteed. The probability of chance separation can be kept low by keeping the number of descriptors used at any one time to below one-third the number of observations (*44*) or, when the class sizes are unequal, to below the number of observations in the smallest class (*45*). In general, the fewer the descriptors used, the more likely that real relationships will be uncovered.

Since the compounds in the mutagen study were diverse and no clear classification scheme was obvious, the analysis was started with simple, easily generated descriptors. If the required classifications are possible by using simple descriptors, then there is little benefit in initially generating complex descriptors. These can be generated if and when they are required.

The simple descriptors alone usually do not achieve very substantial classifications. The definition of "good" classification depends on the study at hand. Certainly 100% correct classification of all the data points is the most desirable result. Allowance must always be made, however, for the outliers and incorrect preclassifications. Also, many of the descriptors are calculated based on approximations or on the properties of isolated substructural fragments that do not consider the bulk of the molecule. A set of such descriptors might not be expected to completely account for all variation in the training set. Few regression studies result in correlation coefficient values of one and PR studies should not always be expected to yield 100% classification.

In the mutagen study the first set of 12 simple fragment and molecular connectivity descriptors achieved 73% correct classification. In view of the complex structures and the variety of functional sites, this is not surprising. It simply means that there was insufficient information within the descriptors to completely describe the mutagen–nonmutagen relationship that was assumed to be present within the training set.

Often many forms of data analysis engage in "exploration" of the data. Exploratory data analysis consists of searching for a suspected relationship among data. In the mutagen study, this meant screening a series of potentially valuable physical properties and structural features that may affect the biological activity. Feature selection is a means of identifying both useful and useless descriptors. There are many methods of feature selection of which variance feature selection is one. There is an extension of the weight-sign method of feature selection and is superior to that technique (*46*). A brief example will illustrate.

In the mutagen study, the descriptor set at this early stage of analysis was incapable of completely separating the training set, but it was capable of separating about 74 of the 105 compounds. The descriptor set did contain some useful information. Before new descriptors were added to this set, the descriptors were screened to see if they were all valuable or if there were extraneous variables that could be eliminated.

As stated previously, in all but severely restricted cases the hyperplane which separates any two classes is not unique; many will serve the same purpose. These different hyperplanes will be defined by different weight vectors that are composed of the weighting coefficients for the descriptors. As the hyperplane shifts position the coefficients which define that plane will change also. Intrinsic variables, those which contribute to the separation of the classes, will be constrained by the class distribution while nonintrinsic variables have fewer constraints and so are free to vary more widely. Variance feature selection is performed by calculating many hyperplanes that separate the separable compounds. The variance of each component of the correspondig weight vectors is a qualitative indicator of the contribution that the corresponding descriptor makes to the separation. Useless variables will generally have much higher variances than those which are useful in the separation. More detailed explanations of the theory and procedure can be found elsewhere (*5*).

A conceptually simpler, but less elegant and much more time-consuming method with which to feature select is to systematically eliminate one or several descriptors at a time and to apply PR methods to this reduced set.

From this point, a cycle of descriptor generation, PR analysis, and feature selection was followed. In order to determine those features that are important to classification and that have not, as yet, been tried, generation of new descriptors is often aided by the examination of previously misclassified patterns, the current descriptor set, and descriptors previously evaluated and discarded. Once simple descriptors prove insufficient to the task, more complex descriptors must be tried. Often this involves using substructure based descriptors. Substructural descriptors have the advantage of directly coding for functional groups which, of course, determine

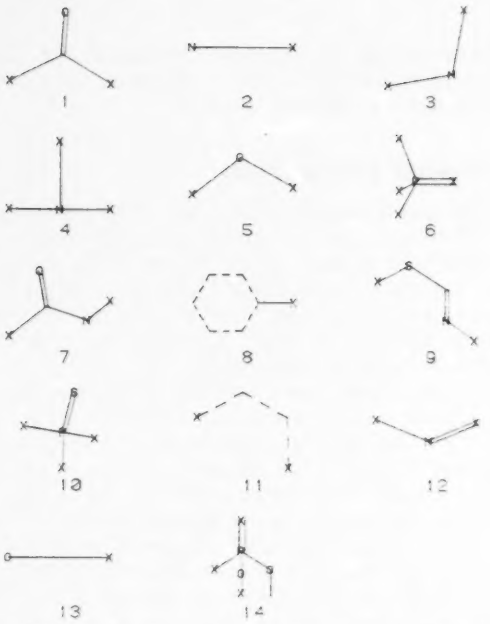
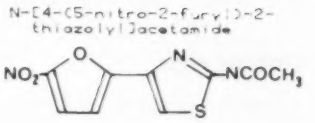


FIGURE 3. Substructures used in descriptor generation for the mutagen study. Hydrogens are suppressed and dashed lines indicate aromatic bonds. X signifies an atom of unspecified type and unspecified connectivity.

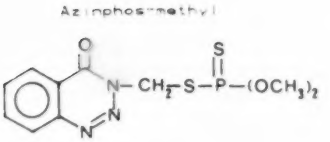
much of a compound's chemistry. Figure 3 shows some of the substructures that were examined in the course of the mutagen study. In order to code for as large a fraction of the compounds as possible within this diverse group, the substructures were kept simple. In a more homogeneous study, more detailed substructures may be of advantage.

The cycle of descriptor generation, PR analysis, and feature selection continued until the addition of further descriptors failed to improve the results appreciably. If the addition of one descriptor allows for the classification of only one additional pattern, then it is unclear whether that descriptor actually adds additional information or whether it simply provides the PR routines with a meaningless additional degree of freedom. Such a descriptor should be critically evaluated before it is used in a study. A similar criterion is also used when removing descriptors from a set. Often a set is reduced until further reduction causes a large increase in the number of misclassified patterns.

During this study a total of 64 descriptors were examined using the PR routines. At no time were more than 30 used in any one set, well within the three-to-one limitation. The study was terminated when a set of 10 descriptors was found which classified 90% of the training set. Eleven compounds were misclassified; five of the mutagens and six of the nonmutagens. This final descriptor set is presented in Table 2. Figure 4 contains sample descriptor values for two of the training set compounds.



1) 1 0
2) 22 59
3) 1 38
4) 0 024
5) 0 0
6) 2 43
7) 20 0
8) 19 0
9) 0 201
10) -0 350



1) 2 0
2) 26 84
3) 1 19
4) 0 046
5) 3 47
6) 0 0
7) 12 0
8) 18 0
9) 0 190
10) 0 0

FIGURE 4. Descriptor values for two compounds in the mutagen training set. Numbering is keyed to the descriptors shown in Table 2.

Table 2. Final descriptor set for the mutagen data set.

		Descriptor
1	Number of sulfur atoms	Fragment
3	Intermediate principal axis	Geometric
2	Number of paths per atom	Complexity - whole molecule
4	Molecular connectivity, valence cluster 5	Complexity - substructural
5	Molecular connectivity, valence path 1 of the immediate environment of a phenyl ring	
6	Molecular connectivity, valence path 1, of the immediate environment around a thiazole fragment (substructure 9)	
7	Sum of the number of paths around an ether oxygen	
8	Sum of the number of paths around doubly and singly bonded nitrogen	
9	del Re sigma charge on a carbonyl carbon	Electronic
10	del Re sigma charge on a secondary amine nitrogen	

Internal Validation

Internal validation serves as a first step in the assessment of the predictive ability of a descriptor set. These procedures are usually performed by forming several subsets of the training set by removing one to many of the observations into separate prediction sets. Discriminants generated from the reduced sets are used to predict the activity of the excluded compounds.

The internal validation procedure must not be thought of as a test of the classification ability that a discriminant would be expected to show with true unknowns. The actual predictive ability of an LDF can be assessed only through verifiable prediction of the activity of compounds that were not involved in the descriptor set development. The descriptor set was generated by analysis of the members of the training set and is dependent on those members. Internal validation results tell us something about the integrity of the data and the descriptor sets but are not a replacement for actual external prediction. Poor internal validation results are, perhaps, more informative than good results. Poor results should cause a reexamination of the descriptor set, the data set, or both. Obviously, if the descriptor set and the reduced training sets can not provide discriminants that will correctly classify the members of the overall training set, then there is little chance of obtaining reliable external prediction results.

Internal consistency checks were performed in the mutagen study by forming 100 subsets of the training set minus the 11 misclassified patterns. The corresponding prediction set for each of the subsets consisted of five mutagens and five nonmutagens which were picked at random from the separable compounds. For the final set, reported in Table 2, the average internal prediction result over the 100 sets was 90%.

External Prediction

The main objective of this study was to develop a set of descriptors and a discriminant which could classify mutagens as defined by the *E. coli* WP2uvrA assay. The descriptor set in Table 2 gave what appears to be good classification results and the internal consistency results were encouraging. The study was ready for the final test—external prediction.

External prediction constitutes extrapolation of the relationship found within the training set. As such, the general rules for extrapolation should apply. A relationship should not be extrapolated beyond the range of the original data. This caution is easily applied to a two-dimensional plot but extension to higher dimensional cases is difficult. Some index or indices of similarity are needed to assess the similarity of an external compound to the training set compounds. When dealing with structurally homogeneous sets of compounds, extreme cases of dissimilarity are easily identified. Within this heterogeneous set of mutagen compounds, however, even extreme cases could be missed by visual examination.

The most obvious place to start is to check to see whether the descriptors even code for the external compound. If a compound has no features similar to the training set, then there is no basis for external prediction. This may seem obvious, but it should be noted that a pattern filled with zeros is a perfectly valid point in

Table 3. Prediction set compounds

No.	Compounds	No.	Compounds
	Mutagens		Nonmutagens
1	2-Nitrofluorene	20	2-Aminofluorene
2	*N*-Acetyl-*N'*-nitrosotryptophan methyl ester	21	1-Naphthylisothiocyanate
3	*N*-Acetyl-*N'*-nitrosotryptophan	22	*p*-Bromoaniline
4	Dimethoate	23	Anethole
5	Epichlorohydrin	24	Estragole
6	Diethyl sulfate	25	Cinnamyl alcohol
7	2,4-Dinitrophenylhydrazine	26	Cinnamaldehyde
8	*N*-Ethyl-*N*-(2-methylallyl)-*N*-nitrosamine	27	Carbon disulfide
9	*N*-Propyl-*N*-(2-methylallyl)-*N*-nitrosamine	28	Safrole
10	4-Nitropyridine-*N*-oxide	29	1,2-Benzanthracene
11	8-Nitroquinoline	30	9,10-Benzanthracene
12	Thioquanine	31	1,1,1-Trichloroethane
13	Psoralen	32	2-(2',4'-Diaminophenoxy)ethanol
14	8-Methylpsoralen	33	Butter Yellow
15	Angelicin	34	Diethylstilbestrol
16	4,5'-Dimethylangelicin	35	4,4'-Methylene-bis-(2-chloroaniline)
17	6-Bromo-4,4-dimethylcyclohexanone	36	Tetrachlorvinphos
18	Dimetridazole	37	Azoxybenzene
19	Ipronidazole	38	Ethionine
		39	Nalidixic acid
		40	Hexamethylphosphoramide

d-dimensional hyperspace, and carelessness may allow such "undescribed" compounds to enter into consideration. The mean number of descriptors coded per compound in the training set was five and this was used as a cutoff value for the external prediction compounds.

Similarity was also maintained by comparing the Mahalanobis distances of the prediction compounds to those of the training set compounds. The Mahalanobis distance of a point is that point's distance from the center of the data cluster in which it is contained. A requirement for external prediction was that the point lay well within the data cluster that was formed by the training set. Forty structurally dissimilar compounds meeting these requirements were obtained from the literature and are listed in Table 3.

In order to perform the predictions, an optimum discriminant was generated from the training set by centering the hyperplane between the classes. This was done by giving the hyperplane its greatest possible "thickness" by using what has been called a deadzone. Since there is no unique weight vector for a particular problem, and there is no clear criterion for choosing one of the many discriminants, centering in this way seems to be a reasonable choice.

This final vector was used to predict the activity of the external compounds. Prediction results were poor—50% overall, 45% for the mutagens and 57% for the nonmutagens. This is approximately what one would expect from random chance.

Difficulties

While the training set results looked promising, carefully executed external prediction failed. How can we account for these seemingly contradictory results? Can such results be avoided? In the remainder of this paper we will endeavor to answer these questions by examining both the data and the methodology. We will highlight several difficulties that occurred in this study and we will show how such problems can be avoided.

The structurally diverse nature of the compounds placed the study under suspicion. Analysis of the heterogeneous structures raises many questions. First, can the data really be subdivided into two distinct classes? If not, how does the presence of subclasses affect PR analysis? Second, if a two-class problem is justifiable, can descriptors be found that will adequately describe the activity of such diverse structures? Third, if the answer to either of the above questions is "no," then what are the chances of arriving at fortuitous results that are equivalent to the results obtained for the training set? We will address each of these questions in turn.

Mechanism of Mutagenesis

We have assumed that mutagenesis is a two-class problem, that mutagens constitute a distinct class apart from nonmutagens. Is this a realistic assumption? A closer look at the mechanism of mutation may answer this question.

Mutations are defined as heritable changes in the DNA; a mutagen changes the original sequence of bases in the DNA. Mutagens cause these changes by several mechanisms. Base analogs are often incorporated directly into the DNA and flat molecules, such as acridines, are known to intercalate between the base pairs. In most cases, however, mutagens react electrophilically with nucleophiles within the DNA (47). These reactions are known to occur at most of DNA's mucleophilic sites. Mutagens react at the O-6, N-2, N-3, N-7, and C-8 positions of guanine. Adenine is attacked at its N-6, N-1, N-3, and N-7 positions. Cytosine reacts at its O-2, N-3, and N-4 atoms, and thymine at its O-2, N-3, and O-4 sites. The phosphodiester backbone also has sites that react with mutagens. Some mutagens show selectivity toward different sites (48). These multiple sites of reaction may, in themselves, fragment the two classes into many.

Metabolism creates an additional complication. For many compounds, metabolic conversion precedes their mutagenic activity. Often the compound that is being evaluated, the promutagen, is not itself a mutagen; it must first be converted by metabolism into the final reactive moiety, the ultimate mutagen. Often the ultimate mutagen bears little resemblance to the promutagen. This alkylating agent, the N-nitrosamines, are a good example of this. The N-nitrosamine promutagens are converted to simple carbonium ion ultimate mutagens before a mutagenic effect is seen (49).

Eight compounds in the study presented here required external activation before they elicited a mutagenic response. They were nonmutagenic without the treatment. In additon, *E. coli* cells are known to contain several enzymes that participate in metabolic activation of mutagens (50,52). The compounds that are coded using the descriptors may have been very different from those which actually damaged the DNA. Metabolism alone may create many classes where two were assumed.

Damage to DNA within *E. coli* cells can be repaired by several repair mechanisms (52,53). This also complicates the problem. Different types of damage will be repaired by different mechanisms and to different extents. In fact, error-prone repair has been implicated in some forms of mutagenesis (54).

Now the problem can be seen in a different light: mutagenesis is a complex problem consisting of many stages. The assumption of two classes for the very different structures within the data set becomes difficult to maintain. Each structural type, each functional group, will have its own chemistry governed by its intramolecular environment. A compound's transport into a cell, its metabolism and remetabolism by a variety of different mechanisms, the site of its interaction with DNA, and the repair processes initiated by the damage that it causes, will all be determined by the chemistry of its overall structure and its functional groups. Within the training set, even the more homologous series of compounds may have different chemistry. Of the seven styrene oxides in the training set, six are mutagenic while one is nonmutagenic. The m-chloro and m-methyl de-

FIGURE 5. A hypothetical two-space for mutagens and nonmutagens. The two classes are distinct but cannot be separated by a single linear discriminant function.

FIGURE 6. An optimal single linear decision surface for the hypothetical problem shown in Fig. 5.

rivatives are mutagenic; however, the *m*-methoxy derivative is not. It seems that each structural class might require a separate study. The data space may consits of not two classes, but many.

By necessity, this review of the mechanisms of mutagenesis has been brief. Many references are available for those interested in a more complete description of this subject (*48*,*55*).

Effects of Diversity

Each additional subclass that is included in a particular study lends greater complexity to the data set and makes the separation into distinct "active" or "inactive" classes more difficult, especially if the presence or degree of subgrouping is unknown. Figure 5 illustrates a simple, hypothetical, two-dimensional two-class problem. The mutagens cluster strongly in two separate regions of the two-dimensional space, but they are still well separated from the strongly clustered nonmutagens. In this simple case the imaginary descriptors are quite capable of separating the mutagens from the nonmutagens. Such a descriptor set is the goal of feature selection procedures.

The problem arises with discriminant development. This example is best separated by two lines as shown in Figure 5. However, if only two classes are assumed then the LDF developing methods would attempt to separate the data into two classes with a single hypersurface. An optimal single hypersurface appears in Figure 6. While it represents the best classification results that can be achieved with a single discriminant, these results are still far less desirable than the complete separation that this descriptor space ideally allows. These results would reflect on the descriptor set which would probably be abandoned in favor of further descriptor development.

This hypothetical case is very simple. While cluster analysis can sometimes uncover clustering in the higher dimensional space that is usually required for SAR studies, such clustering is often difficult to detect due to diffuse and oddly shaped classes, outliers, and misclassified data. The difficulties also increase when one of the clusters is not well represented. Such a case is shown in Figure 7. Clustering analysis may not identify the smaller cluster and the compounds it contains may skew the developing surface as shown in Figure 7 or be misclassified in favor of the larger subgrouping.

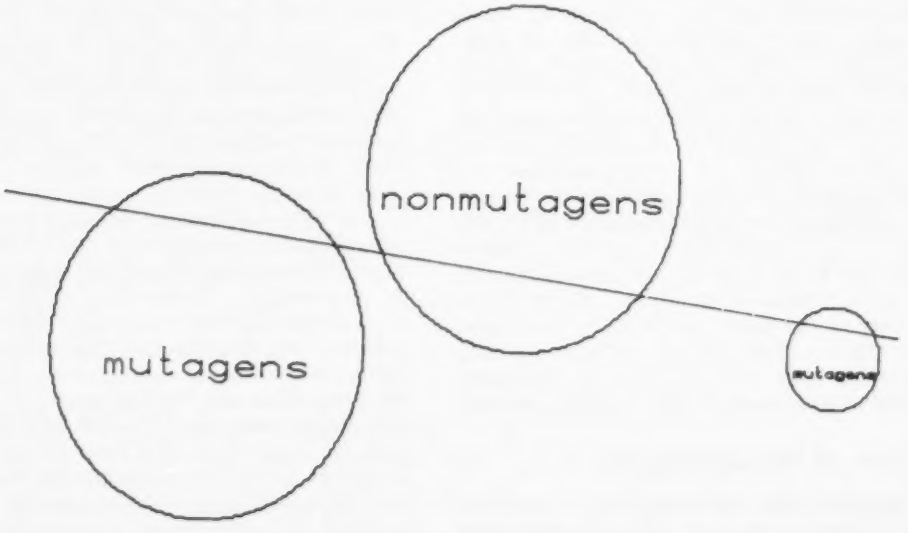


FIGURE 7. Another hypothetical example showing two mutagen clusters but one is poorly represented. A single skewed linear decision surface is shown.

There is no reason for attempting to develop more than one discriminant given what appears to be a two-class problem, however. Doing so would be an unsound practice that could quickly lead to trivial solutions of no real value or predictive capability. The solution lies in the identification of subgroupings of data.

Coding Heterogeneous Structures

The data space that was provided to the PR routines for the mutagen study may have been complicated by many subgroups. Perhaps no hyperplane can be found that will adequately separate the mutagens from the nonmutagens. Also, if each subclass is not well represented or if the number of subclasses is unknown, cluster analysis may be of little value. The problem is not well formulated.

But, what will occur if we insist on maintaining the two class mutagen–nonmutagen problem? Could the complexity of the space have been due to a poor choice of descriptors? After all, the descriptors determine the structure of the data space. If we maintain that a mutagen is a mutagen and a nonmutagen is a nonmutagen could we, with greater effort, still find a "core" of mutagenicity through some "ultimate" descriptor set?

This brings us to the second question regarding heterogeneous data sets: Can such sets be properly coded? If so, then how many descriptors are needed? We noted previously that there is a limited number of degrees of freedom. Can the data be adequately described and still stay within these limits?

Examination of other PR–SAR studies may help to answer these questions. Henry and Jurs reported such a study involving some multiply substituted 9-anilinoacridines (*11*) (Fig. 8). They developed a set of 18 descriptors that was capable of 94% classification of 213 compounds and achieved up to 86% correct external prediction. These external results indicate that the descriptor set codes information pertinent to the activity. It should be noted here that a study of any homologous series already has a good deal of information implicitly coded in the identical structural backbone that does not need to be explicitly coded by descriptors. The coding that is required to explain the activity among the compounds consists of sites and types of substitution, a task whose difficulty should not be underestimated. One can also begin a study by assuming a common mechanism, although this may not always be the case. We will return to this study shortly.

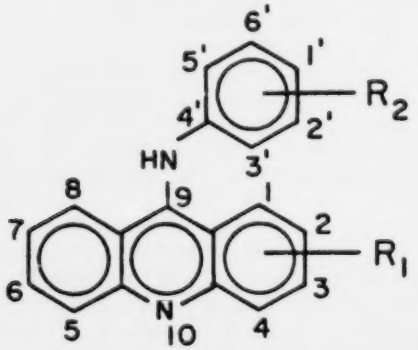


FIGURE 8. The 9-anilinoacridine backbone of the compounds used in the study by Henry and Jurs. Multiple substituents were at the 2,3,4,5,6,1',2', and 3' positions.

Another PR study was performed by Ham and Jurs (*56*). The goal of this study was to classify olfactory stimulants; musks and nonmusks. The compounds involved in this study contained six different structural backbones. The training set of 140 compounds were 100% correctly classified by seven descriptors. If the receptor mediated theory of olfaction is correct, then a common thread runs among these compounds. Activity in such a case can be explained in terms of the "fit" of a particular compound into such a receptor. The descriptor set had to code not for the heterogeneous structures but for the single, or possibly several similar, receptor sites.

What features should be coded in the mutagen study? There is no similar backbone and no receptor site. The activity depends not on some general whole-molecule property but on the reactivity of the functional groups within each molecule. No common thread runs through the mutagens. But, while each structural type and functional group will have its own chemistry, can we still attack the problem by coding each type individually?

With its structural backbone information already implicitly coded, the 9-anilinoacridine study required 18 descriptors to classify 94% of its training set. The mutagens contained at least 12 different structural groups. How many descriptors, how much information, is needed to code for each? Most of these compounds were smaller than the anilinoacridines and so may not require 18 descriptors. Could they be coded by four or five? If so, then the study would require 50 to 60 descriptors. This is far above the 35 degrees of freedom which this study is allowed in order to avoid fortuitous separations. Also, if each of the structural classes is coded separately, some descriptors will code for only the one or two compounds which are contained within that class. This is a situation that should be avoided in PR studies and, in fact, in almost any numerical analysis. Another serious problem of such an approach is insufficient representation of the structural classes. Can any structural class that may contain four, five, or ten sites of substitution be adequately described using a handful of compounds? The answer is probably "no."

When the biological activity of a class of compounds depends on specific chemical reactions and not on whole molecule properties, common intermediates, or specific receptor sites then heterogeneous data sets can not be adequately represented.

Chance Factors in Pattern Recognition

The poor external prediction results may have been due to unrealistic classification, "confusion" of the data

space, insufficient degrees of freedom, improper representation, or a combination of all of these. Why, then, were the classification results for the training set so good?

The answer may lie in three separate but related phenomena: expected random classification due to the number of degrees of freedom, random classification based on fortuitous feature selection results, and classification due to dependence on trivial differences in the training set. We will address each of these phenomena in turn.

Previously it was noted that the number of variables used in a study must be kept below approximately one-third the number of patterns in the smallest class. Above these limits the probability of achieving random 100% correct classifications becomes unacceptably large. In fact, if the number of dimensions used equals the number of observations, 100% classification is assured. On the other end of the scale, with no descriptors at all, 50% of the patterns can be correctly classified by simply assigning all of the patterns to one class. This, of course, is a minimum value for equal class sizes. For an unequal class distribution of 70/30, 70% correct classification can immediately be achieved by assigning all of the patterns to the more populous class. A success rate of 58% can be achieved by classifying the patterns with the same probabilities as the class memberships if they are split 70/30.

These are the well-known extremes but often the middle of the range is forgotten. Each additional dimension will increase the percentage of random, apparently correct, classifications. For classes of equal size these classifications start at 50% with no descriptors at all and increase until 100% classification is guaranteed at the number of descriptors equal to the number of observations. We conducted a series of studies in order to better understand this phenomenon. We found that for 100 patterns which were split equally between two classes, 10 random descriptors can be expected to support such artifactual classifications of about 75%.

These artifactual classifications depend on several factors including the number of observations, the number of descriptors, the relative class sizes, and the overall distribution of the data. The details of these studies are presented elsewhere (*57*,*58*).

This does not mean that classification results using relatively large numbers of descriptors will always be fortuitous. It does mean that a large set of descriptors will always apparently achieve good classification regardless of whether the descriptor set actually can describe the SAR which is occurring. There are two ways to evaluate the "realness" of classification results. First, as will be shown elsewhere, the percentage classification due to artifacts is easily determined and can be compared to the value obtained within an actual study. The difference between the two values will serve as an indicator of the actual information contained within the descriptors. Second, the artifactual classifications are not based on any relationship and so will not extrapolate. A weight vector that does not explain the SAR within the data will have no real predictive ability.

The feature selection methods used may also contribute to fortuitous classifications. The mutagen study was conducted by searching through 64 different descriptors for an optimum set. Such exploratory analysis is common in data analysis and certain criteria determine whether a set is considered optimal or useful. An important index in regression analysis is the correlation coefficient R. A high value of R indicates that the descriptors account for a large percentage of the variation in the dependent variable, the biological activity in SAR studies. An important indicator in PR analysis is the percentage of correctly classified patterns. Such classification results indicate that the descriptors contain information pertinent to the class differences. In searching a pool of numbers for a set which will satisfy these criteria, what are the chances of finding such a set by chance; a set with no SAR significance?

Topliss and Edwards investigated this phenomenon for regression analysis (*59*). They showed that the chance correlation becomes a very real possibility as the number of variables searched becomes large in relation to the number of observations included in a study.
We have investigated similar phenomena for PR analysis and have found that classification results for a training set can be improved due to chance if large numbers of descriptors are searched. A Monte Carlo study of 100 randomly classified patterns in eight dimensions achieved 86% correct classification after searching 32 descriptors. There was no information in the descriptors and no relationship between the patterns. The feature selection procedure merely "chose" those numbers which improved the classification results, the criterion for a valuable descriptor set. This was only one example, and in similar trials the same classification results required a total pool ranging from the 32 to 60 descriptors. Complete separation was possible, but required a larger pool of variables. The details of these studies will be presented elsewhere (*60*).

This is not to imply that classification results achieved after searching a large pool will always be fortuitous or that only small numbers of descriptors should be tried. It does mean, however, that the descriptors used should be carefully generated and selected for chemical significance. "Shotgun" approaches consisting of the generation of copious quantities of descriptors in hopes of finding a useful set should be avoided. This problem also highlights the need for an external prediction set. Fortuitous classifications for the training set will not extrapolate, and external prediction results will be poor.

These two studies were performed using random, continuous, completely uncorrelated variables that are not always equivalent to "real" variables. The differences between the data types are currently under investigation but we feel that much the same trends prevail. We are also investigating the effects this has on the allowed degrees of freedom for a particular study.

The third difficulty is the possibility of classifying the training set due wholly or in part to class differences that are completely dependent on the training set; fea-

tures that may or may not affect the biological activity. An extreme example will illustrate this. Suppose some benzoic acid esters are classified as either quick to hydrolyze or slow to hydrolyze. Additionally suppose that these esters were formed from a wide range of alcohols and that all of the rings are multiple substituted by a variety of substituents. Now, perhaps due to random chance, synthetic requirements, or availability of the esters, suppose all the "quickly" hydrolyzing compounds have a *p*-methyl group. The presence of the *p*-methyl group is a descriptor that provides 100% correct classification of the training set. The *p*-methyl group is almost certainly not the only factor affecting the rate of hydrolysis, however. The number of compounds and the range and quantity of substitution may serve to occlude this defect in homogeneity if the compounds are not closely examined. This is a simple example, but the concept is real. Any feature that is skewed toward one class or the other but is not really involved in the activity can skew the PR results. Other, less severe but still unequal, partitioning of values between the classes are possible and will have similar effects. While the chemical insignificance of this "ideal" descriptor is obvious, other "ideal" descriptors which appear more chemically significant or are complex or are imbedded within multidimensional sets of descriptors may be more subtle.

Some such subtle cases may have been present in the mutagen data set. Close examination of the compounds showed that the bulk of compounds containing phosphorus were mutagenic, the bulk containing sulfur were nonmutagenic, nitrofurans were predominantly mutagenic, and phenyl rings occurred more frequently within the nonmutagenic compounds. While these features probably play a part in the chemistry of mutagenesis, their unequal distribution between the two classes could cause training set dependencies. In order to demonstrate this, indicator variables were generated that coded only for the presence or absence of the substructural features. These four descriptors alone were capable of correctly classifying 75% of the training set compounds. Six additional descriptors of meaningless random numbers were then added for a total of ten, the number in the final descriptor set in Table 2. PR analysis with this manufactured descriptor set achieved the same classification results as with the real descriptor set, 90% correct. Two of these features are found in the final descriptor set and the other two figured prominently in other sets. What would have been the result if more nonmutagens with phenyl rings had been screened by the Gene-Tox report, or more nonmutagenic sulfur compounds? Would the results have been different if none of the mutagenic nitrofurans had been included in this study? In such a case, we would be justified, within the confines of the training set, in saying that nitrofurans were nonmutagenic, a statement that certainly is not true.

These indicator variable results are not equivalent to the actual results because different patterns were misclassified in each case. As such, it may not be the best example, but it should serve to illustrate the problem. Differences between the classes that are not related to the differences in activity can lead to erroneous results. Once again, the phrase, "correlation does not imply causation," rings true. External prediction may be useful in evaluating these effects, also.

Of course, in many cases, such trivial differences may not be obvious. An awareness of the problem, however, is the first step toward its solution. An investigator should be familiar with the compounds in the data set and should deal as closely as possible with chemistry and actual mechanism of activity. The descriptors should be evaluated critically to determine the extent to which trivial training dependent features are coded. Contrarily, if there is a feature that is significant to the SAR and is always present only in one class, then it would be foolish not to include it in an analysis. However, the investigator should be aware of the dangers involved in indiscriminant use of such variables.

Conclusion

In this paper we have presented an outline of the steps taken in performing an SAR study using PR. We have also presented some problems that are inherent in this type of analysis. Pattern recognition is a set of numerical techniques and as such it suffers from many of the same weaknesses that other numerical techniques have. An awareness of these weaknesses is needed in order to avoid their effects. Data set development must be given serious consideration. The true diversity of a study is defined by the mechanism of the activity. The classes used must be determined by the chemistry involved in their activity and not by convenient man-made classifications. In order to avoid fortuitous results, the number of descriptors employed in the descriptor sets should be kept to a minimum. They should be chemically plausible and generated with the mechanism of activity in mind, if it is known. Screening large numbers of descriptors or using large numbers in any one descriptor set may lead to spurious results. The investigator should also be aware of trivial, training set-dependent class differences that could lead to real classifications which cannot be extrapolated to a wider range of compounds. Fortuitous classifications will be revealed by doing external prediction. Such predictions should be done sparingly if the number of compounds available is small. The same compounds should not be consistently used for such predictions. Doing so actually involves them in descriptor set development and they lose their valuable unbiased attributes.

Pattern recognition using linear discriminant functions should not be confused with regression and should not be expected to provide quantitative prediction results. The magnitude of the weight vector components should be interpreted only with care for they are dependent on the overall descriptor set and are not unique.

We have endeavored to supply some guidance but, unfortunately, there are no concrete guidelines for conducting a study. The number of compounds required, the apparent diversity, and the descriptors that will be

useful are all dependent on the study at hand.

Development of the methodology and definition of its limitations continue. Chance factors, similarity measures, and the development of new descriptors are active concerns.

While the study that we have presented has pedagogical value, it should not discourage the use of the techniques presented here. We believe that computer-aided screening of mutagens is possible and we continue to work toward that goal. While it seems clear that mutagenesis is a complex process that will not easily be explained, many of the difficulties which we encountered in this study can be avoided by studies involving mechanistically similar subgroups of mutagens. The information gained from this simpler approach will not only be valuable alone but might also be applied toward attacking more complex problems.

This research was supported by the U.S. Environmental Protection Agency under Cooperative Research Agreement Number CR 807531. The contents do not necessarily reflect the views of the Agency and no official endorsement should be inferred. The PRIME 750 minicomputer used was provided with partial financial support of the National Science Foundation.

REFERENCES

1. Tou, J. T., and Gonzalez, R. C. Pattern Recognition Principles. Addison-Wesley, Reading, MA. 1974.
2. Kowalski, B. R., and Bender, C. F. Pattern recognition. II. Linear and nonlinear methods for displaying chemical data. J. Am. Chem. Soc. 95: 686–693 (1973).
3. Varmuza, K. Pattern Recognition in Chemistry. Springer-Verlag, New York, (1980).
4. Wold, S. Pattern recognition by disjoint principal component models. Pattern Recognition 8: 127–133 (1976).
5. Stuper, A. J., Brugger, W. E., and Jurs, P. C. Computer Assisted Studies of Chemical Structure and Biological Function. Wiley-Interscience, New York, 1979.
6. Harper, A. M., Duewer, D. L., Kowalski, B. R., and Fasching, N. L. ARTHUR and experimental data analysis: the heuristic use of a polyalgorithm. In: Chemometrics: Theory and Application. A.C.S. Symp. Ser. 52 (B. Kowalski, Ed.), American Chemical Society, Washington, DC, 1977, pp. 14–52.
7. Jurs, P. C., Chou, J. T., and Yuan, M. Studies of chemical structure-biological activity relations using pattern recognition. In: Computer Assisted Drug Design, A.C.S. Symp. Ser. 112 (C. E. Olson and R. E. Christopherson, Eds.), American Chemical Society, Washington, DC, 1979, p. 103.
8. Yuan, M., and Jurs, P. C. Computer assisted structure-activity studies of chemical carcinogens. Polycyclic aromatic hydrocarbons. Toxicol. Appl. Pharmacol. 52: 294–312 (1980).
9. Yuta, K., and Jurs, P. C. Computer assisted structure-activity studies of chemical carcinogens. Aromatic amines. J. Med. Chem. 24: 241–251 (1981).
10. Jurs, P. C., Ham, C. L., and Brugger, W. E. Computer assisted studies of chemical structure and olfactory quality using pattern recognition techniques. In: Odor Quality and Chemical Structure (H. R. Moskowitz and C. B. Warren, Eds.), A.C.S. Symp. Ser. 148, American Chemical Society, Washington, DC, pp. 143–160.
11. Henry, D. R., Jurs, P. C., and Denny, W. A. Studies of structure-antitumor activity relations of 9-anilinoacridines using pattern recognition . J. Med. Chem. 25: 899–908 (1982).
12. Chou, J. T., and Jurs, P. C. Computer assisted structure-activity studies of chemical carcinogens. An *N*-nitroso compound data set. J. Med. Chem. 22: 792–797 (1979).
13. Rose, S. L., and Jurs, P. C. Computer assisted studies of structure-activity relationships of *N*-nitroso compounds using pattern recognition. J. Med. Chem. 25: 769–776 (1982).
14. Henry, D. R., and Block, J. H. Pattern recognition of steroids using fragment molecular connectivity. J. Pharm. Sci. 69: 1030–1034 (1980).
15. Dunn, W. J., III, World, S., and Martin, Y. C. Structure-activity study of β-adrenergic agents using the SIMCA method of pattern recognition. J. Med. Chem. 21: 922–930 (1978).
16. Dunn, W. J., III, and Wold, S. The use of SIMCA pattern recognition in predicting the carcinogenicity of potential environmental pollutants. In: Structure-Activity Correlation as a Predictive Tool in Toxicology (L. Goldberg, Ed.), Hemisphere Publ., New York, 1983, pp. 141–150.
17. Norden, B., Edlund, U., and Wold, S. Carcinogenicity of polycyclic aromatic hydrocarbons studied by SIMCA pattern recognition. Acta. Chem. Scand. B32: 602–608 (1978).
18. Jurs, P. C. Studies of relationships between molecular structure and biological activity by pattern recognition methods. In: Structure-Activity Correlation as a Predictive Tool in Toxicology (L. Goldberg, Ed.), Hemisphere Publ., New York, 1983, pp. 93–110.
19. Jurs, P. C., Noor Hasan, M., Henry, D. R., Stouch, T. R., and Whalen-Pederson, E. K. Computer-assisted studies of molecular structure and carcinogenic activity. Fund. Appl. Toxicol. 3: 343–349 (1983).
20. Nilsson, N. J. Learning Machines. Mc-Graw Hill, New York, 1965.
21. Moriguchi, I., Komatsu, K., and Matsushita, Y. Adaptive least-squares method applied to structure–activity correlation of hypotensive *N*-alkyl-*N'*-cyano-*N''*-pyridylguanidines. J. Med. Chem 23: 20–26 (1980).
22. Pietrantonio, L., and Jurs, P. C. Iterative least squares development of discriminant functions for spectroscopic data analysis by pattern recognition. Pattern Recognition 4: 391–400 (1972).
23. Isenhour, T. L., and Jurs, P. C. Some chemical applications of machine intelligence. Anal. Chem. 43: 20A–35A (1971).
24. Kowalski, B. R., and Bender, C. F. Pattern recognition. A powerful approach to interpreting chemical data. J. Am. Chem. Soc. 94: 5632–5639 (1972).
25. Jurs, P. C., and Isenhour, T. L. Chemical Applications of Pattern Recognition. John Wiley and Sons, New York, 1975.
26. Draper, N. R., and Smith, H. Applied Regression Analysis, 2nd Ed. Wiley-Interscience, New York, 1981.
27. Brusick, D. J., Simmon, V. F., Rosenkrantz, H. S., Ray, V. A., and Stafford, R. S. An evaluation of the *Escherichia coli* WP2 and *WP2uvrA* reverse mutation assay. Mutat. Res. 76: 169–190 (1980).
28. Waters, M. D., and Auletta, A. The Gene-Tox Program: genetic activity evaluation. J. Chem. Inf. Comput. Sci. 21: 35–38 (1981).
29. Bru[illegible], and Jurs, P. C. Molecular structure input program[illegible]orage cathode ray tube terminal. Anal. Chem. 47: 781–
30. Brugge[illegible], E., and Jurs, P. C. UDRAW (Program No. 300), Quantum Chemistry Program Exchange, Department of Chemistry, Indiana University, Bloomington, IN 47401.
31. Hansch, C., and Leo, A. Substituent Constants for Correlation Analysis in Chemistry and Biology. John Wiley and Sons, New York, 1979.
32. Chou, J. T., and Jurs, P. C. Computer assisted computation of partition coefficients from molecular structures using fragment constants. J. Chem. Inf. Comput. Sci. 19: 172–178 (1979).
33. Vogel, A. I. Textbook of Practical Organic Chemistry. Longman, New York, 1978, p. 1035.
34. Pearlman, R. S. Molecular surface areas and volumes and their use in structure activity relationships. In: Physical Chemical Properties of Drugs (S. H. Yalkowsky, A. A. Sinkula, and S. C. Valvani, Eds.), Marcel Dekker, New York, 1980, pp. 321–347.
35. Bondi, A. van der Waals volume and radii. J. Phys. Chem 68: 441–451 (1964).
36. Goldstein, H. Classical Mechanics. Addison-Wesley, Reading, MA, 1950, pp. 144–156.
37. Del Re, G. A Simple MO-LCAD method for the calculation of charge distributions in saturated organic molecules. J. Chem. Soc. 1958: 4031–4040 (1958).
38. Randic, M., Brissey, G. M., Spencer, R. B., and Wilkins, C. L. Search for all self-avoiding paths for molecular graphs. Comput. Chem. 3: 5–13 (1979).

39. Kier, L. B., and Hall, L. H. Molecular Connectivity in Chemistry and Drug Research. Academic Press, New York, 1976.
40. Cramer, R. D., III, Redl, G., and Berkoff, C. E. Substructural analysis. A novel approach to the problem of drug design. J. Med. Chem. 17: 533–535 (1974).
41. Hodes, L. Computer-aided selection of compounds for antitumor screening: validation of a statistical-heuristic method. J. Chem. Inf. Comput. Sci. 21: 128–132 (1981).
42. Craig, P. N. Mathematical models for toxicity evaluation. In: Annual Reports in Medicinal Chemistry, Vol. 18 (H. J. Hess, Ed.), Academic Press, New York, 1983, pp. 303–306.
43. Tinker, J. F. Relating bacterial mutagenic activity to chemical structure. In: Structure-Activity Correlation as a Predictive Tool in Toxicology (L. Goldberg, Ed.), Hemisphere Publ., New York, 1983, pp. 207–219.
44. Stuper, A. J. and Jurs, P. C. Reliability of nonparametric linear classifiers. J. Chem. Inf. Comput. Sci. 16: 238–241 (1976).
45. Whalen-Pederson, E. K., and Jurs, P. C. The probability of dichotomization of a binary linear classifier as a function of training set distribution. J. Chem. Inf. Comput. Sci. 19: 264–266 (1979).
46. Zander, G. S., Stuper, A. J., and Jurs, P. C. Nonparametric feature selection in pattern recognition applied to chemical problems. Anal. Chem. 47: 1085–1093 (1975).
47. Miller, J. A., and Miller, E. C. Ultimate chemical carcinogens as reactive mutagenic electrophiles. In: Origins of Human Cancer (H. Hiatt, J. Watson and J. Winsten, Eds.), Cold Spring Harbor Laboratory, 1977, pp. 605–627.
48. Felkner, I. C., Ed. Microbial Testers. Marcel Dekker, New York, 1981.
49. Magee, P. N. Evidence for the formation of electrophilic metabolites for *N*-nitroso compounds. In: Origins of Human Cancer (H. Hiatt, J. Watson and J. Winsten, Eds.), Cold Spring Harbor Laboratory, 1977, pp. 629–637.
50. Childs, J. J., Nakajima, C., and Clayson, D. B. The metabolism of 1-phenylazo-2-naphthol in the rat with reference to the action of the intestinal flora. Biochem. Pharm. 16: 1555–1561 (1967).
51. McCalla, D. R. Metabolic activation of nitro heterocyclic compounds in bacteria and mammalian cells. In: Short Term Tests for Chemical Carcinogens (H. F. Stich and R. H. C. San, Eds.), Springer-Verlag, New York, 1981, pp. 36–47.
52. Smith, K. C. Multiple pathways of DNA repair in bacteria and their roles in mutagenesis. Photochem. Photobiol. 28: 121–129 (1978).
53. Walker, G. C., Elledge, S. J., Perry, K. L., Bagg, A., and kenyon, C. J. Regulation and function of cellular gene products involved in UV and chemical mutagenesis in *E. coli*. In: Induced Mutagenesis (C. W. Lawrence, Ed.), Plenum Press, New York, 1983, pp. 181–202.
54. Kimball, R. F. The relation of repair phenomena to mutation induction in bacteria. Mutat. Res. 55: 85–120 (1978).
55. Hiatt, H. H., Watson, J. D., and Winsten, J. A., Eds. Origins of Human Cancer. B. Cold Spring Harbor Laboratory, 1977.
56. Ham, C. L. Computer-assisted studies of molecular structure and olfactory quality using pattern recognition. Ph.D. Thesis, Department of Chemistry, Pennsylvania State University, 1983.
57. Stouch, T. R., and Jurs, P. C. Monte Carlo studies of the classifications made by nonparametric linear discriminant functions. J. Chem. Inf. Comput. Sci. 25: 45–50 (1985).
58. Stouch, T. R., and Jurs, P. C. Monte Carlo studies of the classifications made by nonparametric linear discriminant functions. II. Effects of nonideal data. J. Chem. Inf. Comput. Sci., in press.
59. Topliss, J. G., and Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships. J. Med. Chem. 22: 1238–1244 (1979).
60. Stouch, T. R., and Jurs, P. C. Monte Carlo studies of the classifications made by nonparametric linear discriminant functions. III. Effects of screening large pools of variables. In preparation.

This article documents research that was conducted under the auspices of the U.S. Environmental Protection Agency (USEPA). As an arm of the government, the USEPA has a paid up, non-exclusive, irrevocable, worldwide license to reproduce, prepare derivative works, and distribute to the public all of the material contained herein, and to grant others acting on its behalf the right to do so. The USEPA grants to the National Technical Information Service, Department of Commerce, the right to act as its agent in the distribution of this material.